Overview of the Cross-lingual Expert Search (CriES) Pilot Challenge

Philipp Sorg[1], Philipp Cimiano[2], Antje Schultz[3], and Sergej Sizov[3]

[1] Institute AIFB, Karlsruhe Institute of Technology
philipp.sorg@kit.edu
[2] Cognitive Interaction Technology, Center of Excellence (CITEC), Bielefeld University
cimiano@cit-ec.uni-bielefeld.de
[3] Information Systems & Semantic Web, University of Koblenz
{antjeschultz|sizov}@uni-koblenz.de

Abstract. This paper provides an overview of the cross-lingual expert search pilot challenge as part of the cross-lingual expert search (CriES) workshop collocated with the CLEF 2010 conference. We present a detailed description of the dataset used in the challenge. This dataset is a subset of an official crawl of Yahoo! Answers published in the context of the Yahoo! Webscope program. Further we describe the selection process of the 60 multilingual topics used in the challenge. The Gold Standard for these topics was created by human assessors who evaluated pooled results of submitted runs. We present data showing that the experts relevant for our chosen topics indeed speak different languages. This corroborates the fact that we need to design retrieval systems that build on a cross-lingual notion of relevance for the expert retrieval task. Finally we summarize the results of the four groups that participated in this challenge using standard evaluation measures. Additionally we also analyze the overlap of retrieved experts in the submitted runs.

1 Introduction

The CriES workshop — Cross-lingual Expert Search: Bridging CLIR and Social Media — addresses the problem of multilingual expert search in social media environments. The main topics are multilingual expert retrieval methods, social media analysis with respect to expert search, selection of datasets and evaluation of expert search results.

In this paper we describe the pilot challenge as part of the CriES workshop. This includes a detailed description of the dataset, the selection process for the topics used in the challenge and the evaluation methodology including relevance assessment. We also present an overview of the results submitted by the participating groups.

Motivation. Online communities generate major economic value and form pivotal parts of corporate expertise management, marketing, product support, CRM,

product innovation and advertising. In many cases, large-scale online communities are multilingual by nature (e.g. developer networks, corporate knowledge bases, blogospheres, Web 2.0 portals). Nowadays, novel solutions are required to deal with both the complexity of large-scale social networks and the complexity of multilingual user behavior.

At the same time, it becomes more and more important to efficiently identify and connect the right experts for a given task across locations, organizational units and languages. The key objective of the lab is to consider the problem of multilingual retrieval in the novel setting of modern social media leveraging the expertise of individual users.

Pilot Challenge Topic and Goals. We instantiate the problem setting by an expert finding task, i.e. our goal is to identify the expertise of online community members and to provide expert suggestions for solving new problems, questions, or help requests in multilingual social media. In many cases, expert users in online communities are multilingual, i.e. they participate in discussions in several languages. Frequently, the actual expertise of the user is language-independent, so he/she could provide meaningful assistance and support to questions and requests stated in any of the known languages. The combined analysis of multilingual user contributions (e.g. answers or postings from the past) together with mining of his social environment (e.g. interaction with other community members in the past, contact/favorite lists, etc.) may provide better indications that the user has the necessary expertise for addressing the request irrespective of the language. The key research challenges addressed by the expert finding task can be summarized as follows:

- User characterization: the use of multilingual evidence of social media for building expert profiles;
- Community analysis: mining of social relationships in collaborative environments for multilingual retrieval scenarios;
- User-centric recommender algorithms: development of retrieval and recommendation algorithms that allow for similarity search and ranked retrieval of expert users in online communities (in contrast to more common document retrieval tasks).

2 Dataset

We used the dataset from the Yahoo! Answers portal introduced by Surdeanu et al. [3].[4] Yahoo! Answers is currently the biggest community QA portal. According to Google ad planner statistics[5] the portal attracts 97M unique visitors and 1.1B page views per month.[6]

[4] This dataset is provided by the Yahoo! Research Webscope program (see `http://research.yahoo.com/`) under the following ID: *L6. Yahoo! Answers Comprehensive Questions and Answers (version 1.0)*

[5] `http://www.google.com/adplanner/`

[6] Statistics from 2010/05/17

The dataset published by Yahoo! contains 4.5M questions with 35.9M answers. For each question one answer is marked as *best answer*. In the portal, the best answer is determined by either the user who submitted the question or via other users ratings. The dataset contains IDs of authors of questions and best answers, whereas authors of non-best answers are anonymous. Questions are organized into categories which form a category taxonomy.

The dataset used in the CriES pilot challenge is a subset of the Yahoo! Answers Webscope dataset, considering only questions and answers in the topic fields defined by the following three top level categories including their sub categories: "Computer & Internet", "Health" and "Science & Mathematics". As our approach is targeted at the expert retrieval problem, by choosing very "technical categories" our goal was to yield a dataset with a high number of technical questions requiring domain expertise to be answered. As many questions in the dataset serve the only purpose of diversion, it was important to find categories were the share of such questions is low.

In the challenge, 4 languages are considered: English, German, French and Spanish. As category names are language-specific, questions and answers from categories corresponding to the selected categories in other languages are also included, e.g. "Gesundheit" (German), "Santé" (French) and "Salud" (Spanish) that correspond to the "Health" category.

The selected dataset consists of 780,193 questions, each question having exactly one best answer. The answers were posted by 169,819 different users, i.e. potential experts in our task. The answer count per expert follows a power log distribution, i.e. 54% of the experts posted one answer, 93% 10 or less, 96% 20 or less. 410 experts published answers in more than one language. These multilingual experts posted 8,976 answers, which shows that they are active users in the portal. The language of questions and answers are distributed over languages as shown in the following table:

		Language share			
Category	*Questions*	*EN*	*DE*	*FR*	*ES*
Comp. & Internet	317,074	89%	1%	3%	6%
Health	294,944	95%	1%	2%	2%
Science & Math.	185,994	91%	1%	2%	6%

2.1 Topic Selection

The topics we use in the challenge consist of 15 questions in each language (60 topics in total). As these topics are questions posted by users of the portal, they express a real information need. Considering the multilingual dimension of our expert retrieval scenario, we defined the following criteria for topic selection to ensure that the selected topics are indeed applicable for our task:

- *International domain.* People from other countries should be able to answer the question. In particular, answering the question should not require knowledge that is specific to a geographic region, country or culture. Examples:

```
   Pro: Why doesn't an optical mouse work on
        a glass table?
Contra: Why is it so foggy in San Francisco?
```

– *Expertise questions.* As the goal of our system is to find experts in the domain of the question, all questions should require domain expertise to answer them. This excludes for example questions that ask for opinions or do not expect an answer at all. Examples:

```
   Pro: What is a blog?
Contra: What is the best podcast to subscribe
        to?
```

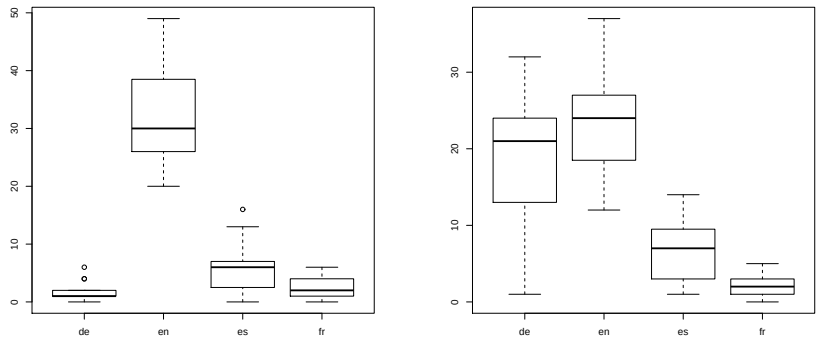
We performed the following steps to select the topics:

1. Selection of 100 random questions per language from the dataset (total of 400 candidate topics).
2. Manual assessment of each candidate topic by three human assessors. They were instructed to check the fulfillment of the criteria defined above.
3. For each question the *language coverage* was computed. The language coverage tries to quantify how much potentially relevant experts are contained in the dataset for each topic and for each of the different languages. The language coverage was calculated by translating a topic into the different languages (using Google Translate) and then using a standard IR system to retrieve all the expert profiles that contain at least one of the terms in the translated query. Topics were assigned high language coverage if they matched an average number of experts in all of the languages. In this way we ensure that the topics are well covered in the different languages under consideration but do not match too many experts profiles in each language. This is important for our multilingual task as we intend to find experts in different languages.
4. Candidate questions are sorted first by the manual assessment and then by language coverage. The top 15 questions in each language were selected as topics.

3 Relvance Assessment

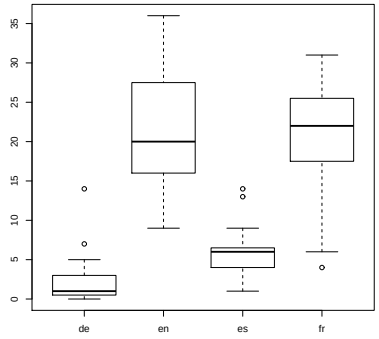
We used result pooling for the evaluation of the retrieval results of the participating groups. For each pooled run, the top 10 experts were pooled and evaluated. The assessment of experts was based on expert profiles. Assessors received topics and the complete profile of experts, consisting of all answers posted by the expert in question. Based on this information they assigned topic-expert tuples to the following relevance classes:

2 Expert is likely able to answer.
1 Expert may be able to answer.
0 Expert is probably not able to answer.

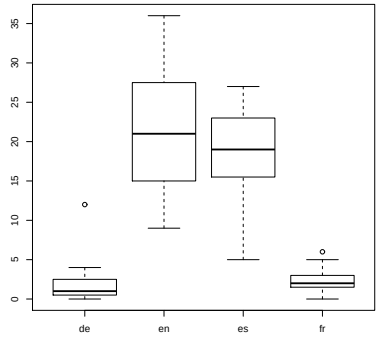


(a) English Topics

(b) German Topics



(c) French Topics



(d) Spanish Topics

Fig. 1. Distribution of relevant users for topics in different languages. Users are classified to languages based on their answers submitted to the Yahoo! Answers portal and for each user class a separate distribution is visualized.

The assessors were instructed to only use evidence in the dataset for their judgments. It is assumed that experts expressed all their knowledge in the answer history and will not have expertise about other topics, unless it can be inferred from existing answers.

Overall, 6 assessors evaluated 7,515 pairs of topics and expert profiles. The distribution of relevant users for the topics in the four different languages is presented in Figure 1. In order to visualize the multilingual nature of the task we also classified relevant users to languages using their answers in the dataset. The distribution of relevant users for the topics in the four languages is shown separately for each user group. The analysis of the relevant user distribution shows that for all topics the main share of relevant users publish answers either in the topic language or in English. This motivates the cross-language expert retrieval task we consider, as mono-lingual retrieval in the topic language or cross-lingual retrieval from the topic language to English do clearly not suffice. The number of relevant experts posting in a different language than the topic language or English constitute a small share. However the percentage is large enough — for example Spanish experts for German topics — in order not to consider these experts.

4 Results

Baseline. In addition to the submitted runs, we defined a standard IR baseline: BM25+Z-Score. This baseline uses language specific indexes of expert text profiles. These profiles consist of all former answers of each expert in a specific language. Topics are translated to each language using Google Translate and the BM25 model [1] is used to get language specific results. Using the Z-Score normalization [2], the final scores for each expert for a specific topic are obtained by aggregation.

Evaluation of Submitted Runs. Four different groups participated in the pilot challenge. Results based on the relevance assessment of the top 10 retrieved experts are presented in Figure 1. In addition to the submitted runs we also present results for the baseline define above. We use two different evaluation measures: Precision at cutoff level 10 (P@10) and Mean Reciprocal Rank (MRR). The best runs achieved promising retrieval results with P@10 of .62 (strict assessment, iftene_run2) and .87 (lenient assessment, herzig_3-boe-07-02-01-q01m). Both runs significantly improve the baseline that achieves P@10 of .19 (strict assessment) and .39 (lenient assessments). Precision / Recall curves for each run are presented in Figure 2 using strict assessment and in Figure 3 using lenient assessment.

Overlap of Retrieved Experts. The overlap of retrieved experts between runs is presented in Table 2. Comparing any combination of two runs, the presented numbers correspond to the count of retrieved experts for each topic that are not retrieved by both runs.

	Strict		*Lenient*	
Run Id	*P@10*	*MRR*	*P@10*	*MRR*
iftene_run2	.62	.84	.83	.94
iftene_run0	.52	.80	.82	.94
herzig_3-boe-07-02-01-q01m	.49	.76	.87	.93
herzig_1-boe-06-03-01-q01m	.48	.77	.86	.94
iftene_run1	.47	.77	.77	.93
herzig_2-boe-06-03-01-q01	.35	.65	.61	.74
leveling_DCUa	.09	.14	.40	.51
leveling_DCUq	.08	.16	.42	.54
bastings	.07	.15	.25	.43
BM25 + Z-Score	.19	.40	.39	.63

t

Table 1. Results of the runs submitted to the CriES pilot challenge. Precision at cutoff level 10 (P@10) and Mean Reciprocal Rank (MRR) are used as evaluation measures. Results are presented for both strict and lenient assessments.

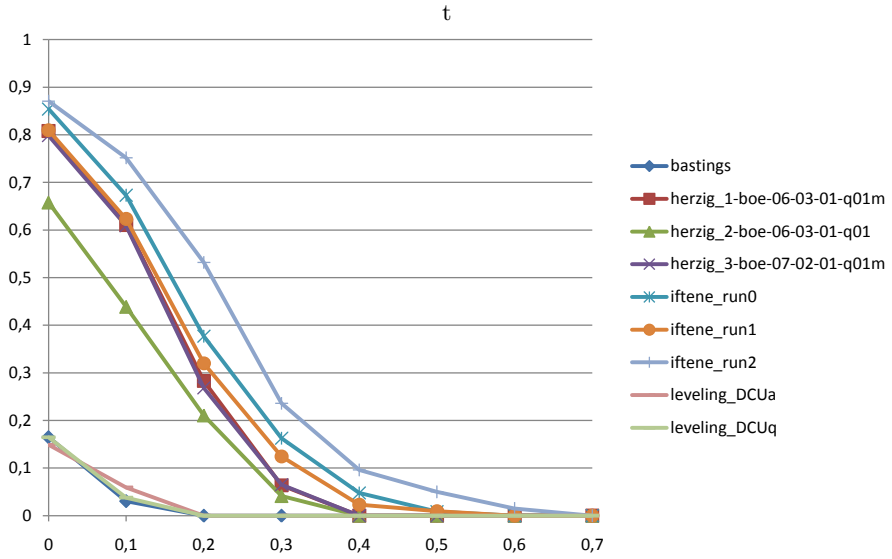


Fig. 2. Precision/Recall Curves based on interpolated Recall (strict assessment).

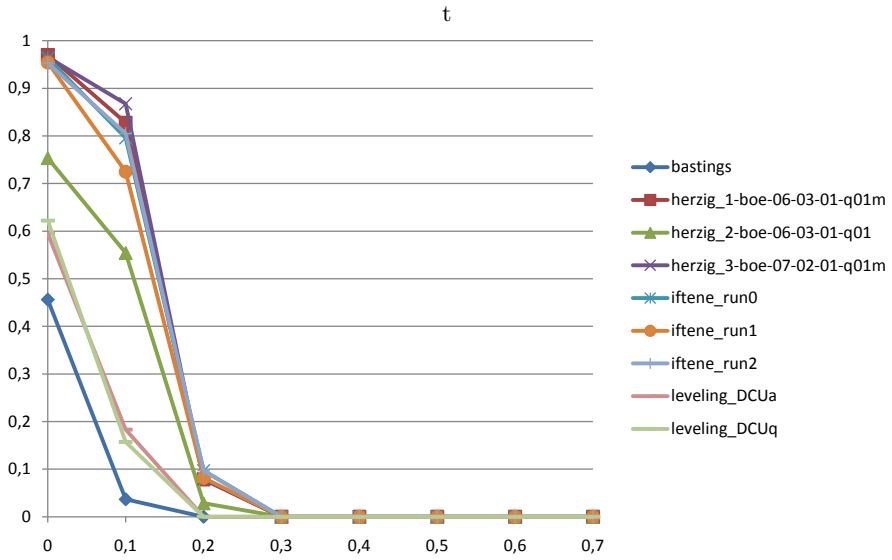


Fig. 3. Precision/Recall Curves based on interpolated Recall (lenient assessment).

t		1	2	3	4	5	6	7	8	9
	1 bastings_submission	0								
	2 herzig_1-boe-06-03-01-q01m	598	0							
	3 herzig_2-boe-06-03-01-q01	598	314	0						
	4 herzig_3-boe-07-02-01-q01m	598	46	324	0					
	5 iftene_run0	599	517	537	517	0				
	6 iftene_run1	598	522	533	522	215	0			
	7 iftene_run2	600	517	533	518	293	352	0		
	8 leveling_DCUa	575	597	599	598	596	589	597	0	
	9 leveling_DCUq	584	596	596	596	597	587	599	528	0

Table 2. Dissimilarity matrix for retrieved experts of the submitted runs. The presented numbers correspond to the count of retrieved experts of each run for all topics that are not retrieved by the compared run.

The presented statistics show that the overlap of retrieved experts across the four groups is very low. Even the best performing runs of two different groups (iftene_run2, herzig_3-boe-07-02-01-q01m) have a small overlap of 14% while having similar values for P@10 and MRR. This shows that the combination of different approaches is an important topic for future work.

Acknowledgments

This work was funded by the Multipla project[7] sponsored by the German Research Foundation (DFG) under grant number 38457858 as well as by the Monnet project[8] funded by the European Commission under FP7.

References

1. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Proceedings of the 17th International Conference on Research and Development in Information Retrieval (SIGIR). pp. 232—241. Springer, Dublin (1994)
2. Savoy, J.: Data fusion for effective european monolingual information retrieval. In: Multilingual Information Access for Text, Speech and Images, pp. 233—244 (2005)
3. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers on large online QA collections. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 719–727. Columbus, Ohio (2008)

[7] http://www.multipla-project.org/
[8] http://www.monnet-project.eu/